**Introduction**

By now, you should understand the concept of profiling data and its practical applications to identify data anomalies. You covered the three primary data profiling operations, column quality, column distribution, and column profile. In this exercise, you can apply your knowledge by using Microsoft Power Query to examine the valid, error, empty, min, max, unique, and distinct values in Excel spreadsheet rows, allowing you to identify the anomalies in the data.

**Case study**

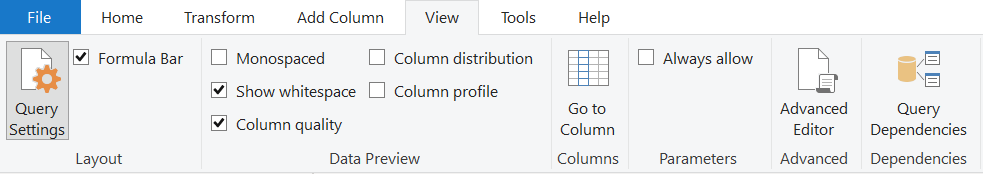
Adventure Works has recently acquired another bicycle business. Adventure Works’ CEO, Jamie Lee, has assigned a task to the sales department to ensure that the product data from the newly acquired company is validated, revised in quality factors, and ready for importing to the current company products. Your manager, Adio Quinn, has requested that you examine the data using Power Query by the factors of profile, quality, and distribution.

Adventure Works has provided an Excel file containing the newly acquired company’s product data called *Other Company Products.xlsx*. The dataset has some empty data in its **ProductKey** column. You also need to assess the distribution of the products by the product categories and detect potential anomalies in the **Price** column. To complete your task successfully, you must import the Excel file, transform the data in Power Query and assess the **Column quality**, **Column distribution**, and **Column profile** options in the **Data Preview** group.

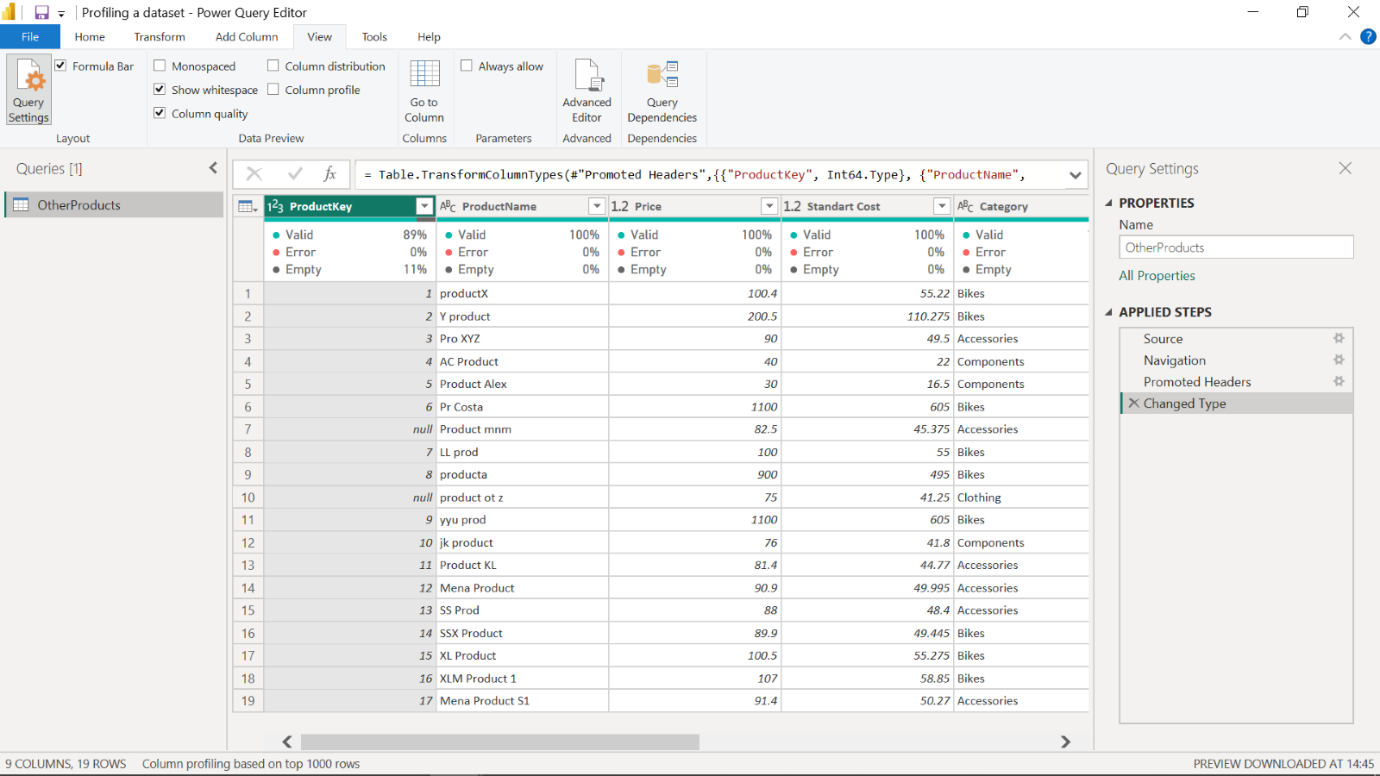
* This exercise aims to help you to understand how to identify data anomalies by profiling data.
* By the end of this exercise, you’ll understand how to profile data in Power Query, and how to identify data anomalies.

### Step 1: Detect empty values in **ProductKey** column

1. There are empty values in **ProductKey** column.
2. To detect empty and invalid values, you need to assess column quality, on the **View** ribbon tab, from inside the **Data Preview** group, and select the **Column Quality** checkbox. **The column quality feature allows you to easily** determine the percentage of valid, error, or empty values found in columns.

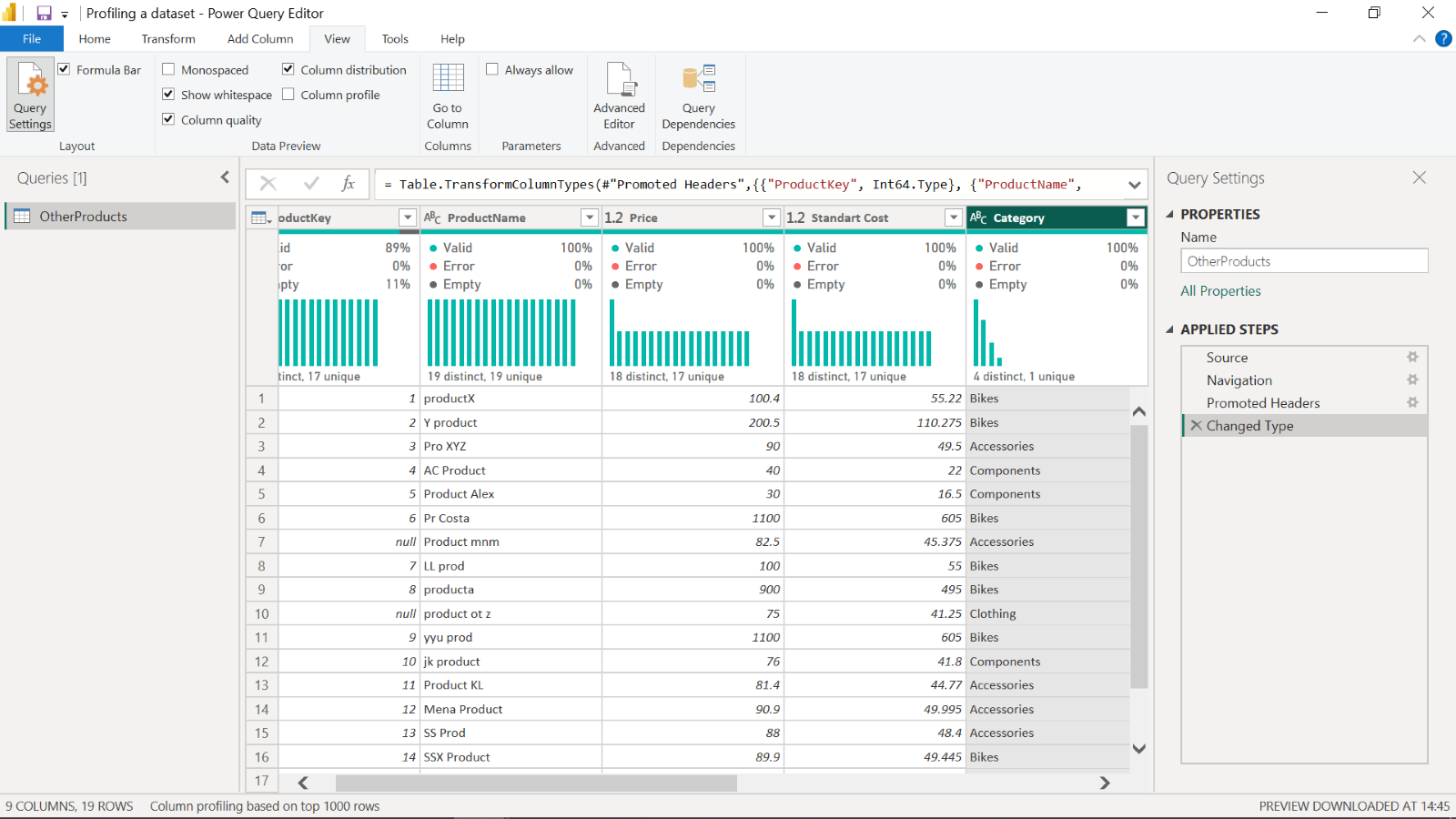


1. Note that 89% of the values are **Valid**, 0% of the values are **Error** and 11% of the values are **Empty** rows in the **ProductKey** column.

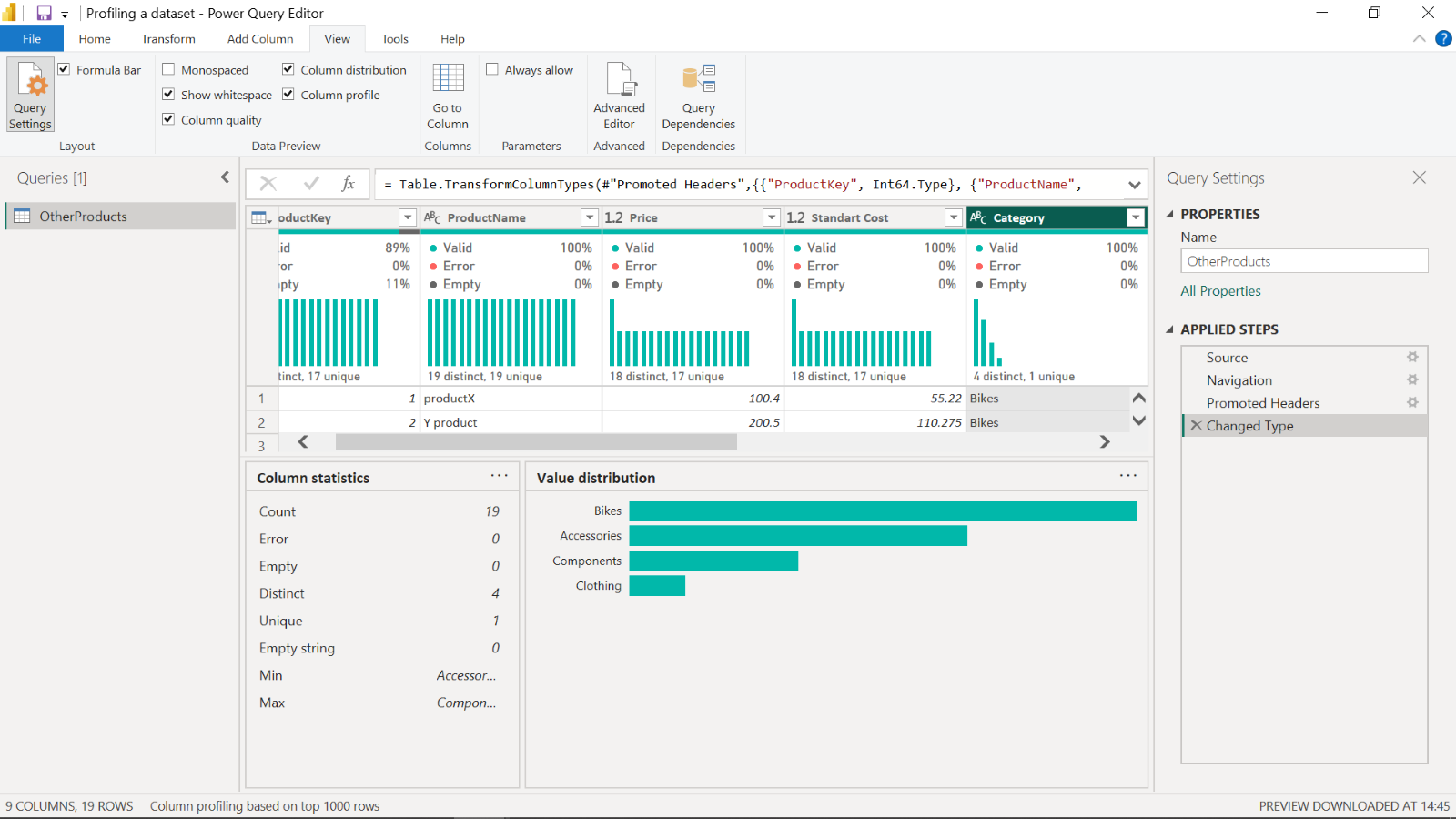


### Step 2: Assess the distribution of product categories

1. There are many categories in the **Product** list, and you need to find out how the data is distributed by the category data.
2. To assess column distribution, on the **View** ribbon tab, from inside the **Data Preview** group, check the **Column Distribution** checkbox. Note that there are 4 **distinct** values and 1 **unique** value.



1. Check the **Column Profile** checkbox, while keeping **Column Distribution** checkbox as checked. Note that there are 9 Bikes, 6 Accessories, 3 Components and 1 Clothing categories.



### Step 5: Detect potential anomalies in the price column

1. You have to assess the **Price** column in **Product** list and you need to find out min, max, mean values and the distribution of the values. To detect potential anomalies and assess column distribution for the **Price** column, on the **View** ribbon tab, from inside the **Data Preview** group, check **Column Profile** while keeping **Column Distribution** checkbox as checked.
2. Note that min value is 30, max value is 1100, and average is 239.13 for the **Price** column.
3. There are 18 distinct and 17 unique values, which means there are 2 products with the same price.
4. When you assess the value distribution, it can be considered as normal distribution and there are some outliers like 30, 40, 900 and 1100. The aim of this assessment is to find a potential anomaly in the **Price** column.

